A GRNN classifier using random generator and application to classifying promoters

Kunho Kim¹, Byungwhan Kim¹, Kyungnam Kim², Jin Han Hong³
¹Sejong University, Electronic Engineering, ²Molecular Biology, ³Macrogen, DNA Chip Division

1. 서론

DNA 칭 정보로부터 질병진단과 신약개발을 위한 유용한 생물학 정보를 추출하기 위한 연구가 활발히 진행되고 있다. 인공신경망은 무항생자 DNA 염기서열성의 전자기기정과 중첩성과 같은 프로모서 확인을 위한 중요한 정보를 예측하고 분류하는데 이용되고 있다 [1-3]. 염기열분석에는 다양한 총류의 신경망이 응용되고 있다. 이 중, 일반화된 회귀신경망 (Generalized regression neural network-GRNN) [4]는 구조가 간단하고 성능최적화를 위한 계산단색도 하나밖에 없어, 많은 응용이 기대되고 있다. 일반적으로, GRNN의 성능을 spread 변수를 실수적으로 결정하여 최적화하며, 패턴의 뉴런은 동일한 최적화된 spread를 가질게 된다. 이에 따라 GRNN의 성능을 개선하는데는 한계가 있었다. 본 연구에서는 난수발생기 (RG)를 이용하여 GRNN의 성능을 향상시킬 수 있는 spread에서 최적화가 되게하는 기법을 제시하고, 본 기법을 프로모터 염기서열의 분석에 적용한다. 고정된 난수발생기와 다양한 분류기를 사용하여, 이 종 결정된 최적의 분류기의 성능을 분류신경망의 예측확정도별 측면에서 평가한다. 분류신경망의 결정계 (Threshold) 값을 변화시켜서 평가한다. 제안된 분류기를 총록의 분류기와 그 성능을 비교평가한다.

2. 본론

2.1 실험데이터 수집


\[ Y = X^T \]  

(1)

\[ W_p = Y \text{ones} \]  

(2)

출력층에는, 단순히 합측의 두 뉴런의 출력을 나누어 예측치를 출력한다. 임의의 입력패턴 x에 대한 예측치는 (3)으로 구해야 한다.
\[
\hat{y}(x) = \sum \frac{\exp[-D(x,x_i)]}{\sum \exp[-D(x,x_i)]} y_i 
\]

여기서 \(x_i\)는 저장된 \(i\)번째의 학습패턴을 지칭하며, \(n\)은 전체 학습데이터의 수를 의미한다. 
(3)에서 함수 \(D\)는
\[
D(x, x_i) = \sum (x - x_i)^2 
\]

여기서 \(p\)는 각 입력패턴을 구성하는 전체 독립변수의 수를 지칭한다. \(x_i\)는 \(x\)와 \(x_i\)의 \(j\)번째 요소의 유사성을 의미한다. 그리고 변수 0은 폭 \(\text{spread}\)을 불러, GRNN의 성능을 결정하는 유일한 학습인자이다. 일반적으로 \(\text{spread}\)는 반복적으로 일정한 범위에서 결정하며, 결정된 \(\text{spread}\)은 그림 1의 편평성을 구현하는 모든 뉴런에 대해서 동일하다. \(\text{spread}\)가 더 작음수록 결정할 때, GRNN의 성능이 향상될 수 있으며, 이와 관련한 연구는 미미한 상황이다.

2.3 분류기의 성능 분석
분류기 성능을 예측정확도와 분류밀도도 측면에서 평가한다. 예측정확도는 (5)로 정의된 RMSE에 의해 계산한다.
\[
\text{RMSE} = \frac{1}{p} \left( \sum_{i=1}^{p} (d_i - \text{out}_i)^2 \right) 
\]

여기서 \(p\)와 \(q\)는 출력층 뉴런의 수와 테스트 페턴의 수를 의미한다. \(d_i\)와 \(\text{out}_i\)는 \(i\)번째 페턴에 대한 \(j\)번째 출력뉴런에 주어지는 실제치와 그 뉴런으로부터의 예측치를 의미한다. 분류밀도도는 입력의 class로 정확히 분류되는 테스트 입력패턴의 수로 정의된다. 분류밀도도는 (6)의 임계값 (threshold)을 기준으로 더욱 세분화하여 평가한다.
\[
d_i < \text{Threshold} 
\]

임계값은 0.6~0.9로 변화시켰으며, spread는 0.4~1.4배에서 0.1간격으로 증가시켜 주었다. 한편, RG-가기 분류기의 경우 주어진 난수 범위에서 200개의 모델을 발생시켰다.

2.3.1 고정 spread를 가진 분류기
주어진 범위의 범주로 분류기를 설계한다. 주어진 spread에서 구성한 분류기의 전체 RMSE가 그림 2에 도시되어 있다.

그림 2. 전체 RMSE의 spread에 따른 변화

여기서 RMSE는 테스트 페턴 전체에 대해서 계산된 값이다. 그림 2에서와 같이 RMSE는 spread의 증가에 따라 거의 선형적으로 증가하고 있다. 그림 2에서, 분류기는 spread가 1.4에서 최적의 RMSE (0.5313)을 가진다.
다음에는 각 spread에서 계산된 RMSE와 (6)을 이용하여 전체 분류밀도도를 임계값의 함수로 계산하였으며, 그 결과가 그림 3에 도시되어 있다.

그림 3. 전체 분류밀도도의 spread에 따른 변화

그림 3에서와 같이, 전체 분류밀도도는 spread의 증가에 따라 갑작스럽게 감소하고 있다. 모든 임계값에 대해서, 분류기는 0.4에서 가장 우수한 전체 분류밀도도를 보인다. 각 임계값에 결정된 전체 분류밀도도를 각 프로모터별로 다시 계산하였으며, 그 결과가 표 1에 정리되어 있다.

표 1. 전체 분류 민감도의 프로모터 별 분류

<table>
<thead>
<tr>
<th>Threshold</th>
<th>OS</th>
<th>AT</th>
<th>EC</th>
<th>ZM</th>
<th>Total CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>14</td>
</tr>
<tr>
<td>0.8</td>
<td>6</td>
<td>0</td>
<td>4</td>
<td>5</td>
<td>15</td>
</tr>
<tr>
<td>0.7</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>0.6</td>
<td>6</td>
<td>0</td>
<td>5</td>
<td>5</td>
<td>16</td>
</tr>
</tbody>
</table>

표 1에서와 같이, 분류기는 OS인 경우에 대해서 가장 우수한 분류성을 보이고 있으며, AT 프로모터인 경우, 임계값에 상관없이 전체 분류도를 유지하고 있다.

표 2. 전체 분류 민감도의 프로모터 별 분류

<table>
<thead>
<tr>
<th>Threshold</th>
<th>OS</th>
<th>AT</th>
<th>EC</th>
<th>ZM</th>
<th>Total CS</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.9</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>8</td>
<td>21</td>
</tr>
<tr>
<td>0.8</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>0.7</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>24</td>
</tr>
<tr>
<td>0.6</td>
<td>1</td>
<td>9</td>
<td>6</td>
<td>8</td>
<td>24</td>
</tr>
</tbody>
</table>

2.3.2 다중 spread를 가지는 분류기와 비교평가

RG를 이용하여 페턴층 뉴런이 다중 값을 갖게 하였으며, 그 성능을 평가한다. 입력의 난수범위에서, 발생된 200개의 각 분류기에 대해서 주어진 임계값에 대한 전체 분류밀도도를 계산하였으며, 그 중 전체 분류밀도도가 가장 큰 분류기에 대한 전체 분류밀도도 값을 그림 4에 도시되어 있다. 그림에서와 같이, 전체 분류밀도도는 그림 3에서와 같이 일반적으로 난수범위의 증가에 따라 감소하고 있다. 모든 임계값에 대해서 난수범위 0.4에 대한 전체 분류밀도도는 가장 우수하였으며, 이는 그림 3에서 인도 결과와 동일하다. 다중에는 각 임계점에 대해서 전체 분류밀도도가 가장 우수한 분류기 성능을 각 프로모터 별로 나타낸다.
로모터 유형별로 세분화하였으며, 그 결과가 표 2에 정리되어 있다. 표 1과 비교할 때, OS의 경우 분류성능이 매우 저하되었다. 그러나 AT의 경우 표 1에서의 결과와는 달리 상당한 분류성능을 보이되 있으며, 이 같은 결과는 매우 주목할 만하다. EC와 ZM에 대해서도 분류 성능이 향상되었다. 평균적으로 prokaryotic 유형의 프로모터에 대한 분류성능이 우수하며, 이는 종류의 분류가에서 판촉된 결과와 동일하다. 표 1과 비교할 때, 전체 분류성능은 임계점 0.9에 대해서 7개 항상되었고, 0.8에 대해서는 9개 항상되었다. 나머지 0.6과 0.7에 대해서는 8개가 항상되어, RC2를 이용한 분류기의 성능이 더 우수하다는 것을 알 수 있다.

다음에는 결정된 최적의 분류기의 성능을 각 프로모터별로 비교 평가한다. 그 결과가 표 3에 정리되어 있다. 표 3은 임계점이 0.9에 대한 결과만을 포함하고 있으며, 이는 다른 임계점에 대해서도 동일한 결과를 얻을 기 때문이다. 표 3에서와 같이, 기존 분류기에 대해서 AT, EC, ZM의 경우 그 RMSE가 향상되었으며, 단지 OS의 경우에 대해서만, 저하된 RMSE를 보이고 있다. 표 3에서와 같이, 일반적으로 RC2 기초의 분류기 설계 방식이 개별적 프로모터의 RMSE를 향상시키는데에도 매우 효과적이었다.

### 표 3. 최적의 분류기성능에 대한 예측정도 비교

<table>
<thead>
<tr>
<th>예측정도 기준</th>
<th>RG-GRNN</th>
<th>GRNN</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>OS</td>
<td>0.675</td>
<td>0.518</td>
<td>-30.3%</td>
</tr>
<tr>
<td>AT</td>
<td>0.419</td>
<td>0.701</td>
<td>40.2%</td>
</tr>
<tr>
<td>EC</td>
<td>0.538</td>
<td>0.577</td>
<td>6.7%</td>
</tr>
<tr>
<td>ZM</td>
<td>0.480</td>
<td>0.574</td>
<td>16.3%</td>
</tr>
</tbody>
</table>

### 3. 결론

RC2 기초의 분류기 설계방식을 제안하였으며, 이를 프로모터 엔지니어링에 적용하여 그 성능을 예측정도와 분류성능을 측면에서 평가하였으며, 분류 성능을 임계점의 함수로 더 세분화하였다. 비교결과, 제안된 분류기는 전체 분류성능을 향상시켰고, 각 프로모터에 대한 분류성능도 분명하게 증진되었다. 또한 최적화된 분류기의 비교에서 알 수 있듯이, 제안된 기법은 전체 RMSE의 개별적 프로모터에 대한 RMSE도 증진시켰다. 이 같은 결과는 제안된 RC2 기초의 분류기 설계방식이 DNA 참테이터의 해석에 효과적으로 응용될 수 있을음을 말한다.

### 감사의 글
본 연구는 IMT 2000 연구비에 의해 지원 받았으며, 한국보건산업진흥원에 감사의 뜻을 드립니다.

참고 문헌